Reactivated visual masks do not disrupt serial recall. A failed Replication of Rey et al. 2018.

Lea Bartsch & Klaus Oberauer

University of Zurich, Cognitive Psychology

Author Note

We thank Atalia Adank and Dawid Strzelczyk for helping with data collection and coding of responses. The data, the preregistration and the analysis scripts can be accessed in the Open Science Framework (https://osf.io/tj4sh/). Correspondence should be addressed to Lea M. Bartsch, Department of Psychology, Cognitive Psychology Unit, University of Zurich, Binzmuehlestrasse 14/22, 8050 Zurich, Switzerland. E-mail: l.bartsch@psychologie.uzh.ch.

Abstract

The process of spontaneous refreshing plays a central role in current models of working memory but is yet to be observed directly. In a recent study, Rey and colleagues (Rey, Versace, & Plancher, 2018) introduced a novel approach to investigate the mechanisms underlying refreshing: They presented tones previously associated with a visual mask during the free time of a complex span task, and found that this impaired memory, presumably because reactivation of the masks disrupts refreshing. Here we aimed to replicate their finding under more controlled settings with more observations per participant. We failed to replicate the previous findings, thereby questioning the robustness of the original effect.

Keywords: refreshing, visual masks, working memory

Reactivated visual masks do not disrupt serial recall

Working memory (WM) is a capacity-limited system for holding the currently most relevant information available for processing (Cowan, 2017). Over the years several processes have been introduced, that supposedly underlie active maintenance in WM, one of which is called *refreshing*: Refreshing has been proposed as a domain-general attentional maintenance process in WM, by which the level of activation of memory traces is restored (Barrouillet et al., 2011; Camos & Barrouillet, 2014, Camos et al., 2018). This process plays a prominent role in the time-based-resource sharing (TBRS) model of WM. According to the TBRS model, refreshing is needed to temporarily maintain transient representations in the face of decay. Refreshing is assumed to rely on a central attentional resource that needs to be shared with other concurrent processing demands such as reading or solving arithmetic problems (Camos, Lagner, & Barrouillet, 2009). So far, it is unclear how refreshing is implemented, whether it is a slow deliberate or a fast spontaneous, scanning-style process, and on which kind of representations refreshing operates (see Camos et al., 2018 for a review). So far, the assumption that people use refreshing spontaneously as a maintenance process relies only on indirect evidence from the trade-off between storage and processing demands: In complex-span tasks, participants encode list items for subsequent serial recall, and in between engage in a secondary processing task. Some studies have varied the *cognitive load* imposed by the processing task, that is, the processing demand per unit time (e.g., Barrouillet et al., 2007; Camos, Mora, & Barrouillet, 2013). As cognitive load is increased, memory decreases. This finding is explained by the TBRS model because higher cognitive load implies more time for decay and less time for refreshing The effect of cognitive load on memory, however, can also be explained in another way: At lower cognitive load, participants have more time to remove representations involved in the

processing task from WM, thereby reducing interference (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012). Therefore, the cognitive-load effect is not compelling evidence for spontaneous refreshing.

To date, only four studies experimentally induced refreshing in tasks to test its effect on memory (visual WM: Souza and Oberauer, 2017a; Souza, Rerko, & Oberauer, 2015; verbal WM and long-term memory: Bartsch, Loaiza, Jäncke, Oberauer, & Lewis-Peacock, 2019; Bartsch, Singmann, & Oberauer, 2018). Yet, it is unclear whether the process induced through retrospective cues as in the aforementioned studies – which was interpreted as reflecting the deliberate refreshing of memory traces – reflects the same mechanism as *spontaneous* refreshing. Moreover, it is not clear whether people even engage in refreshing spontaneously during a WM task (Oberauer, 2019a; Vergauwe et al., 2016).

In a recent study, Rey and colleagues aimed at adding to the merely indirect evidence available to-date that spontaneous refreshing occurs, and that it operates through the reactivation of memory traces. They investigated how the reactivation of an irrelevant trace (a visual pattern mask) prevents attentional refreshing (Rey, Versace, & Plancher, 2018). According to Rey and colleagues, their results indicate that refreshing relies on the reactivation of sensory memory traces.

Rey et al. (2018) conducted three experiments in which they first associated a tone with a visual mask in an initial learning phase. Subsequently they presented the tone at various time points during a complex-span task. The tone was supposed to reactivate the visual mask and thereby impede refreshing, which should have a detrimental effect on recall performance. Experiment 1 presented the tone during encoding of visual stimuli, Experiment 2 presented the tone during the free time of a complex span task (the time refreshing supposedly occurs), and the

third Experiment presented the tone during the distractor task, when refreshing is assumed to be impossible anyway. Compared to the presentation of a tone associated with a control stimulus, the tone associated with a visual mask disrupted oral serial recall when it was presented at encoding (Experiment 1) and during free time (Experiment 2) but not during the distractor task (Experiment 3). The authors concluded that encoding (Experiment 1) and maintenance (Experiment 2) of visual information in working memory during a complex span task is disrupted by the reactivation of another visual memory trace. They interpret these findings – in particular the effect during maintenance in Experiment 2 – as evidence in favour of the refreshing mechanism assumed in the TBRS theory.

Rey and colleagues proposed their method as a means for directly manipulating and thereby investigating the effects of refreshing in working memory. Having a tool to experimentally test the postulated hypotheses about the role and characteristics of refreshing in current memory models would be desirable. Therefore, we wanted to test the robustness of the effects described by Rey et al. (2018). If the findings of Rey and colleagues were robust, their method would allow researchers to investigate how refreshing operates, and answer the many open questions about the process described above and in a recent review (see Camos et al., 2018).

Yet, some aspects of the study of Rey and colleagues raise the question how robust their findings are: First, the experiment only consisted of 16 trials, with 8 trials per mask condition. Given the small numerical difference in mean performance (54.3% in the mask condition compared to 59.2% in the control condition), we believe more trials per participant are needed to ensure that the effect is measured reliably. Further, responses were noted by the experimenter,

which could be subject to experimenter error or even bias (if the experimenter knows which condition is currently running).

Here we aimed to directly replicate the critical Experiment 2 of Rey and colleagues in a more controlled setting including more trials per participant. We preregistered the present study (see https://osf.io/tj4sh).

Methods

Participants

We collected data of 30 participants, a similar sample size to the original study, which included N = 28 participants We had preregistered three exclusion criteria: Participants were to be replaced if: (1) their overall performance across all conditions was close to chance; (b) they did not complete all experimental conditions; or (c) if they did not comply with the instructions to read out aloud the digits presented as distractors or to orally recall the memoranda at recall. Based on these criteria, 6 subjects were replaced, as they did not comply with the instruction to orally recall the memoranda. Only participants whose mother tongue is German, aged between 18-35 years, and reporting normal or corrected-to-normal vision and hearing took part in the experiment. Participants signed an informed consent form prior to the study and were debriefed at the end. The experimental protocol is in accordance with the regulations of the Ethics Committee of the Faculty of Arts and Social Sciences of the University of Zurich.

Materials and Procedure

We used the material provided to us from the authors of the original study, consisting of (1) the twenty-five black-and-white pictures of animals and objects, (2) the original control and visual mask, as well as (3) the high- and low-pitched tones. The labels corresponding to the

pictures were translated into German by the first author, and they did not differ in word length compared to the French labels of the original study (French: M_{length} = 6.67 letters, Geman: M_{length} =6.13 letters), and ranged comparably in length (French: range = [3,12]; German range = [4,12]). The labels can be found in Table A1. The experiment was set up and managed in Matlab with Psychophysics toolbox 3. Equivalent to the original study, the subjects underwent an association phase, followed by the test phase (see Figure 1). Because we drastically increased the number of trials in the test phase from 16 to 64, we introduced a second association phase to "refresh" the association between masks and tones after half of the test-phase trials (i.e. after 32 trials). For each participant, tones were consistently associated with the same visual stimulus throughout the experiment.

Association Phase

The association phase was the same as described before (Rey et al, 2018). After a fixation point centrally displayed for 500 ms and a 100 ms delay, for half of the participants the visual mask was presented simultaneously with the high-pitched tone whereas the control mask was presented together with the low-pitched tone for 500 ms. For the other half of the participants, the association was reversed. During the simultaneous presentation of the visual stimulus and the tone, the participants were asked to indicate the pitch of the tone by button-press ("h" for high pitched, "t" for low-pitched). The tones were presented via headphones. In total the association phase consisted of 30 trials per mask in random order.

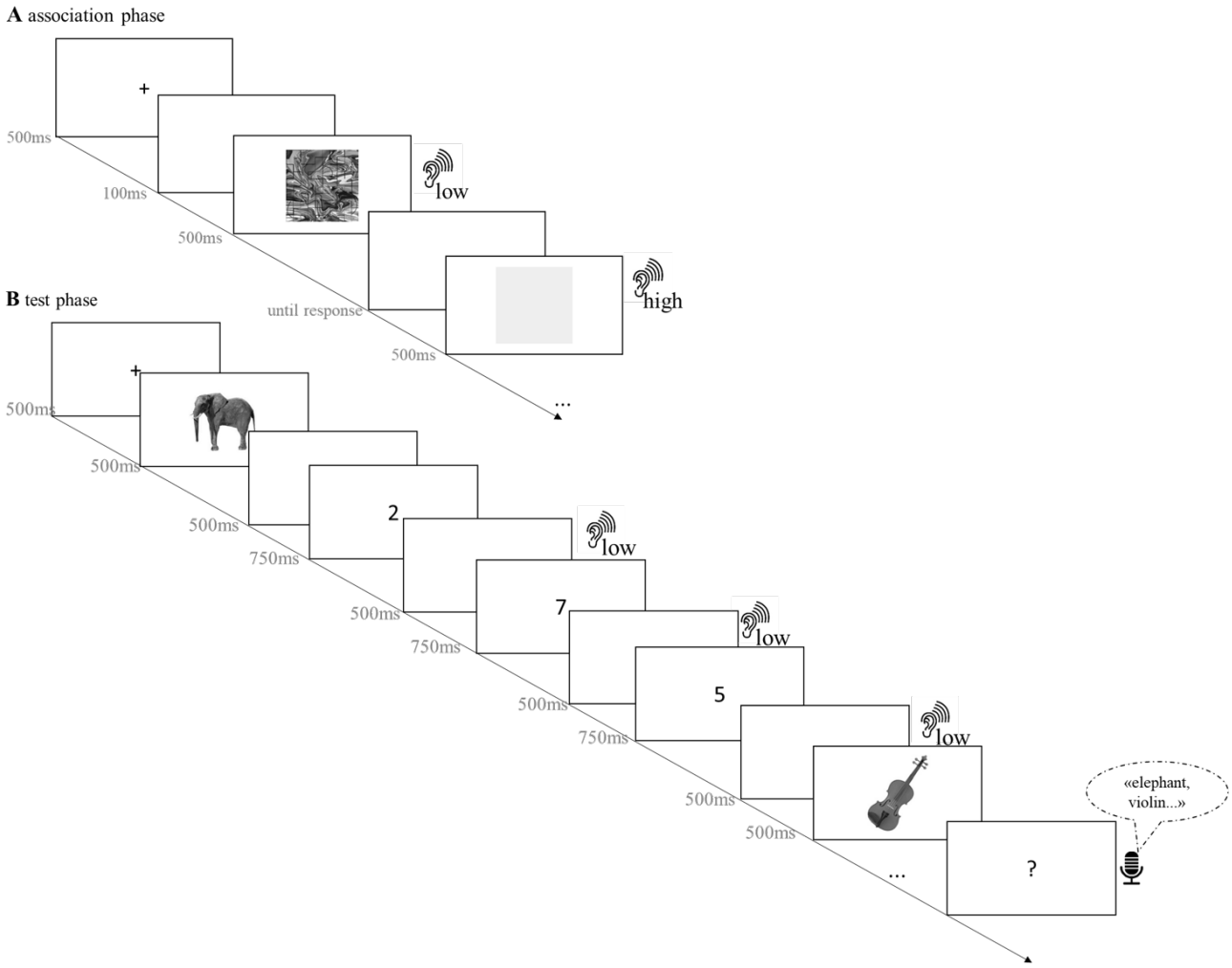

Figure 1 A. The association phase and B. the test phase of the replication study. Materials and timings are identical to Rey et. al (2018).

Test Phase

As in the original study (Rey et al., 2018), trials began with a fixation point centrally displayed for 500 ms, followed by a 100 ms delay, and then five pictures were successively presented for 500 ms per item. Each picture was followed by a series of three digits to be read aloud. These distractors were successively presented for 750 ms each, with an inter-stimulus delay of 500 ms. As in Experiment 2 of Rey et al., tones were presented in the blank-screen

intervals between each distractor. Within a complex-span trial the tone was consistently either associated with the mask, or associated with the control stimulus.

Participants were instructed to name the five pictures as they were presented and to memorize them. At the end of the trial, when a question mark appeared on the screen, they were to recall them by saying their names in their serial order of presentation. Unlike the original study, the participants were recorded throughout the trial to ensure compliance with naming the pictures and reading aloud the digits. We also recorded the oral serial recall of the subjects.

We increased the number of trials to 64, with 32 trials per masking condition (tone associated with visual mask vs. tone associated with control stimulus). Trials of the two masking conditions were randomly intermixed. As in the original study, the same picture was not repeated in consecutive trials. Also, pictures of animals and objects were randomly mixed. Before the test trials subjects performed two practice trials without the presentation of tones.

Results

Analysis

Serial recall accuracy

The recorded responses were coded by a student assistant who was blind to the association of tones to masks of the participants. Matlab was used to enter coded responses to text files. The data were analyzed using Bayesian generalized linear mixed models (BGLMM) implemented in the R package *brms* (Bürkner, 2017, 2018). The dependent variable was the binary outcome of correct (1) or incorrect (0) response per observation in each trial per participant. Correct responses are defined as recalling the target item at the correct serial position. Therefore, we assume a Bernoulli data distribution predicted by a linear model through

a logit link function (i.e., a repeated-measures logistic regression). The fixed-effect was *mask condition* (mask vs. control). Following the recommendation of Barr and colleagues (Barr, Levy, Scheepers, & Tily, 2013; see also Schielzeth & Forstmeier, 2009) we implemented the maximal random-effects structure justified by the design; by-participant random-intercept and by-participant random-slope for condition. In addition, we estimated the correlation among the random-effects parameters.

The regression coefficients were given moderately informative Cauchy priors with scales between 0.3 and 2. These scales were chosen because they define a default prior analogously to that proposed by Rouder et al. (2012) for the General Linear Model and because these priors were recently introduced as default priors for logistic models (Oberauer, 2019b). Specifically, this prior assigns its probability mass approximately equally over those effect sizes on the predictor scale that translate into effects between -0.5 and 0.5 on the p(correct) scale when starting from p(correct)=0.5 as baseline. We used completely non-informative priors for the correlation matrices, so-called LKJ priors with shape parameter 1. We calculated Bayes Factors to estimate the strength of evidence for the null or the alternative hypothesis. In other words, with the BF we can calculate the evidence for the effect of the masking condition (BF_{10}) against an intercept-only model that serves as the null model. Additionally, we can calculate evidence against a difference between the conditions (BF_{01}), where $BF_{01} = (1/BF_{10})$. A BF_{10} larger than 1 gives evidence for an effect, a BF_{10} lower than 1 yields evidence against an effect and hence evidence for the null hypothesis. A BF_{10} of 10 indicates that the data are 10 times more likely under the alternative hypothesis than under the null hypothesis. Usually, BFs > 3 are regarded as providing substantial evidence for one hypothesis over the other. We aimed to report BFs ≥ 10

for or against the alternative hypothesis for the main effect in the model, as a BF ≥ 10 is regarded as strong evidence.

We used an MCMC algorithm (implemented in Stan; Carpenter et al., 2017) that estimates the posteriors by sampling parameter values proportional to the product of prior and likelihood. These samples are generated through 4 independent Markov chains, with 1000 warmup samples each, followed by 50000 samples drawn from the posterior distribution which were retained for analysis. Following Gelman and colleagues (2013), we confirmed that the 4 chains converged to the same posterior distribution by verifying that the $\hat{R}$ statistic – reflecting the ratio of between-chain variance to within-chain variance – was < 1.05 for all parameters, and we visually inspected the chains for convergence. Finally, we used the *bayes_factor* function in the *brms* package, which implements the bridge sampler (Gronau, Singmann, & Wagenmakers, 2017), for computing the BFs.

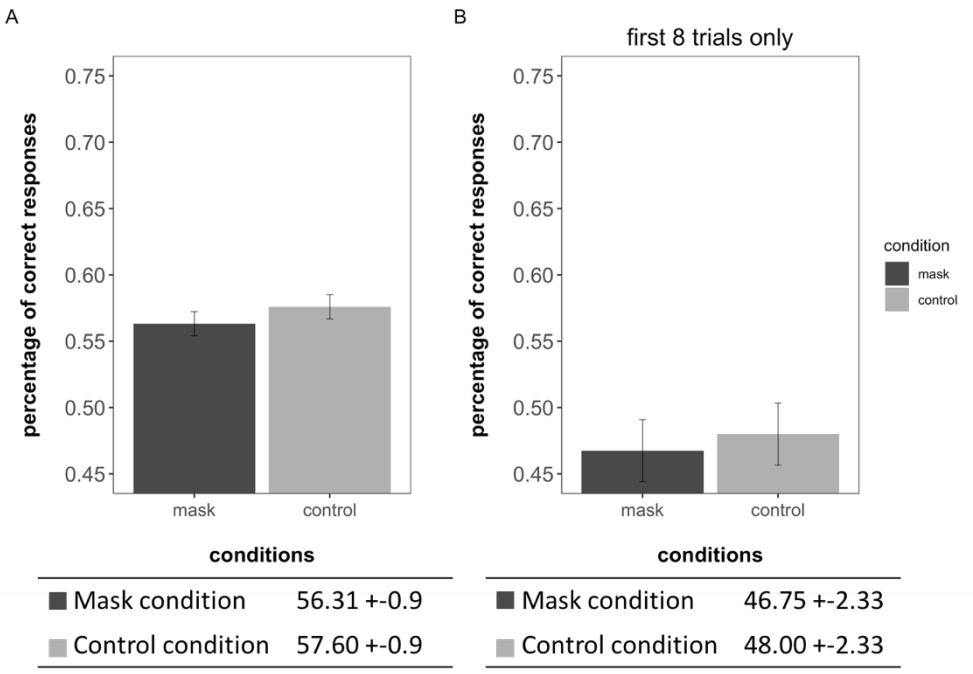

Figure 2 Mean correct responses in percent as a function of the condition (mask vs. control). A: across all the trials and B: across the first 8 trials after each of the association phases only. Error bars represent the standard error.

Association Phase

Participants performed the tone pitch task with a mean correct response rate of 95.22 %. The BF_{01} of 3.01 provides evidence against a difference of correct responses to the high and low pitch sounds.

Test Phase

The serial-recall performance is shown in *Figure 2*A. The comparison of a model including the effect of condition (mask vs. control) to an intercept-only model yielded very strong evidence against a difference between the conditions (BF_{01} = 58.66).

Reanalysis of the original data

Analysis

We obtained the aggregated data of the original study, which was provided by the authors as supplementary material (https://econtent.hogrefe.com/doi/suppl/10.1027/1618-3169/a000414), in order to reanalyze them in a Bayesian framework, similar to how we analyzed the data of the present replication study above.

As the aggregated data does not allow us to implement the same analysis as above, we will describe the changes in the following, then report the results of this analysis of the original data and compare this to results of the same analysis stream with the aggregated data of the present replication study.

We analyzed the original data using Bayesian generalized linear mixed models (BGLMM) implemented in the R package *brms* (Bürkner, 2017, 2018). Here, the dependent variable was the proportion of correctly recalled items per condition and per participant. Therefore, we assume a Gaussian data distribution predicted by a linear model. Again, we calculated Bayes Factors to estimate the strength of evidence for the presence or absence of an effect by comparing the likelihood of the data given several linear models. We therefore specified models to include or omit the fixed-effect of condition (mask vs. control), as well as the by-participant random-slopes for condition.

To further rule out the alternative explanation that our results differed from the original simply due to the larger number of trials following an association phase (here 32 vs. 8 in the original study), we further analyzed only the first 8 trials per condition following the first association phase of our data in the aforementioned framework.

Results

The results of these three separate analyses can be found in Table 1. The BGLMM of the original data of Rey et al. showed anecdotal evidence *against* including the condition factor as fixed effect compared to the Null model ($BF_{nullvsfixed} = 2.33$; this is the BF_{01} corresponding to the Fixed vs. Null model comparison of Table 1). The best model was the Null model, including only participant as a random effect. We also tested the condition effect through the ttest.BF function of the *BayesFactor* package (Morey, Rouder, & Jamil, 2015) to compute the Bayes factor for the Bayesian t-test, corresponding more closely to the frequentist paired-sample t-test of the original study. This analysis showed only anecdotal evidence ($BF_{10} = 2.82$) for an effect of condition in the original study.

Table 1 Results of the Bayesian hierarchical linear mixed model predicting the proportion of correct responses of the replication study, the data of the first 8 trials of each mask after each of the two association phases of the replication study, and the original study.

Model	Fixed effect	Random effect	Comparison	BF_{10}		
				Replication	Replication $T_N = 8$	Original
Full	condition	condition	vs. Fixed	1.94×10^{-3}	1.37×10^{-2}	0.02
Fixed	condition	-	vs. Null	0.03	0.05	0.43
Null	-	-	Vs. Full	**1.97×10^{4}**	**1.60×10^{3}**	**93.85**

Note: All models include participant as a random effect. The model printed in bold is the best model.

The BGLMM of the aggregated data of the present replication study confirmed our initial analyses, showing that there was very strong evidence for the null model, compared to models including condition as fixed effect, and as both, random and fixed effect. This was also true for the case in which we included only the first 8 trials of each condition, thereby simulating the

much shorter experiment of Rey and colleagues. We thereby rule out the alternative explanation that our results differed from the original simply due to the larger number of trials following an association phase (here 32 vs. 8 in the original study). Our results rather confirm that small numbers of observations are subject to overestimating effects that are actually noise.

We extracted the posterior distributions of the effect of condition from the fitted BGLMMs of the three datasets (replication, replication first 8 trials and original data, see Figure 3) to compare the effect sizes of the effect of masking condition. The mode of the posterior provides a point-estimate of the effect size (i.e., the central tendency of the posterior difference). The 95% credibility interval gives the smallest range of effect estimates over which 95% of the posterior probability is concentrated, and as such provides an assessment of the uncertainty of estimation (i.e., the dispersion of the posterior difference). As depicted in Figure 3A, the posterior distributions of the original and replication data overlap strongly, with our replication study showing the narrowest distribution. Therefore, there is no reason to believe that the effect in the original study truly differed from the one in our replication.

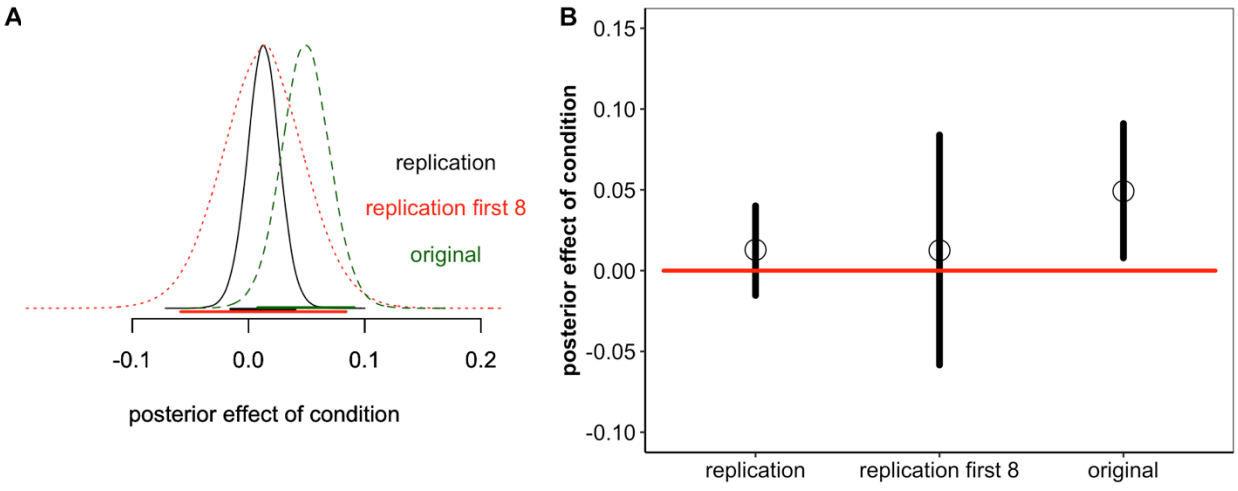

*Figure 3 Posterior distributions of differences of parameters between the conditions of the replication study, the first 8 trials of the replication study and the original. **A** The posterior density and 95% highest*

density intervals which reflect the effect size of any condition difference. ***B*** *The same distributions including their mode with its respective highest density intervals. The red horizontal line characterizes the point of no evidence for an effect.*

Discussion

The present study failed to replicate the findings of Experiment 2 of Rey and colleagues: The presentation of tones associated with visual masks during the free time in a complex span task did not disrupt serial recall performance.

As our Bayesian reanalysis of the original data shows, the evidence for an effect of the mask manipulation was ambiguous to begin with. Nevertheless, one might ask about possible reasons for the slightly different outcomes in the original experiment and our replication. One possible reason for the diverging findings is, that we implemented a larger number of trials following an association phase (here 32 vs. 8 in the original study), which might have decreased the strength of association over the course of the trials. Furthermore, the increase in number of trials lead to an increase in the amount of times stimuli are repeated. Yet, our analysis of only the first 8 trials following the first association phase leads us to the same conclusions as before: that the mask had no effect on serial recall. Instead, what we see is that the effect of the original study is not robust under more controlled conditions, and when the analysis accounts for trial-to-trial noise. The hierarchical BGLMM used here, modelling the correctness of each observation, can separate this trial noise from the true effect.

Another reason for the diverging findings lies in one of the motivations to replicate the study in the first place: The original study did not control for experimenter bias, because the responses of participants were recorded by an experimenter, who may not have been blind to the manipulation. In our replication study, answers were recorded by Matlab and later coded by a condition-blind research assistant.

This replication only focussed on experiment 2 of the original study, as this one was central to the claim made for the process of spontaneous refreshing. Based on our data we therefore cannot draw any conclusions on whether the effect of reinstating masks at encoding would replicate or not.

Although the original findings promised a novel approach to investigate refreshing directly, the present results indicate that the effect of the masks do not hold up in a more controlled setting including more trials per participant. Thereby, a way to experimentally manipulate the effectiveness of spontaneous refreshing is yet to be discovered.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartsch, L. M., Loaiza, V. M., Jäncke, L., Oberauer, K., & Lewis-Peacock, J. A. (2019). Dissociating refreshing and elaboration and their impacts on memory. *NeuroImage*, *199*. https://doi.org/10.1016/j.neuroimage.2019.06.028

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance , and their contributions to long-term memory formation. *Memory & Cognition*, *46*(5). https://doi.org/10.3758/s13421-018-0805-9

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Camos, V., Johnson, M. R., Loaiza, V. M., Portrat, S., Souza, A. S., & Vergauwe, E. (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences*, 1–14. https://doi.org/10.1111/nyas.13616

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). *Stan* : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, 3rd edition*. Chapman & Hall/CRC.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: An R package for estimating normalizing constants. *ArXiv Preprint ArXiv:1710.08162*.

Oberauer, K. (2019a). Is Rehearsal an Effective Maintenance Strategy for Working Memory? *Trends in Cognitive Sciences*, 1–12. https://doi.org/10.1016/j.tics.2019.06.002

Oberauer, K. (2019b). Working Memory Capacity Limits Memory for Bindings. *Journal of Cognition*, *2*(1).

Rey, A. E., Versace, R., & Plancher, G. (2018). When a Reactivated Visual Mask Disrupts Serial Recall Evidence That Refreshing Relies on Memory Traces. *Experimental Psychology*, *65*, 263–271. https://doi.org/https://doi.org/10.1027/1618-3169/a000414

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. https://doi.org/10.1093/beheco/arn145

Vergauwe, E., Hardman, K. O., Rouder, J. N., Roemer, E., McAllaster, S., & Cowan, N. (2016). Searching for serial refreshing in working memory: Using response times to track the content of the focus of attention over time. *Psychonomic Bulletin & Review*, *23*(6), 1818–1824.

Appendix

Table A1. List of labels for the pictures in French (original), German (replication study), and in English for interested readers.

French	German	English
Ane	Esel	Donkey
Chat	Katze	Cat
Chevre	Ziege	Goat
Chouette	Eule	Owl
Elephant	Elefant	Elephant
Grenouille	Frosch	Frog
Lavevaiselle	Spülmaschine	Dishwasher
Loup	Wolf	Wolf
Maracas	Rassel	Rattle
Mouton	Schaf	Sheep
Oie	Gans	Goose
Oiseau	Vogel	Bird
Perroquet	Papagei	Parrot
Poule	Huhn	Chicken
Puma	Puma	Puma
Rasoir	Rasierer	Razor
Reveil	Wecker	Clock
Saxophone	Saxophon	Saxophone
Sifflet	Pfeife	Pipe
Tamtam	Trommel	Drum
Telephone	Telefon	Telephone
Tondeuse	Rasenmäher	Mower
Trompete	Trompete	Trumpet
Violon	Geige	Violin